Title

mirMachine: A One-Stop Shop for Plant miRNA Annotation.

Journal

Journal of Visualized Experiments, 2021(171)

ISSN

1940-087X

Authors

Cagirici, H Busra

Sen, Taner Z

Budak, Hikmet

Publication Date

2021

DOI

10.3791/62430

mirMachine: A One-Stop Shop for Plant miRNA Annotation

H. Busra Cagirici[1], **Taner Z. Sen**[1], **Hikmet Budak**[2]

[1] U.S. Department of Agriculture - Agricultural Research Service, Western Regional Research Center, Crop Improvement and Genetics Research Unit, CA, USA [2] Montana BioAgriculture Inc., Missoula, MT, USA

Corresponding Author

Hikmet Budak
hikmet.budak@icloud.com

Citation

Cagirici, H.B., Sen, T.Z., Budak, H. mirMachine: A One-Stop Shop for Plant miRNA Annotation. *J. Vis. Exp.* (171), e62430, doi:10.3791/62430 (2021).

Date Published

May 1, 2021

DOI

10.3791/62430

URL

jove.com/video/62430

Abstract

Of different types of noncoding RNAs, microRNAs (miRNAs) have arguably been in the spotlight over the last decade. As post-transcriptional regulators of gene expression, miRNAs play key roles in various cellular pathways, including both development and response to a/biotic stress, such as drought and diseases. Having high-quality reference genome sequences enabled identification and annotation of miRNAs in several plant species, where miRNA sequences are highly conserved. As computational miRNA identification and annotation processes are mostly error-prone processes, homology-based predictions increase prediction accuracy. We developed and have improved the miRNA annotation pipeline, SUmir, in the last decade, which has been used for several plant genomes since then.

This study presents a fully automated, new miRNA pipeline, mirMachine (miRNA Machine), by (i) adding an additional filtering step on the secondary structure predictions, (ii) making it fully automated, and (iii) introducing new options to predict either known miRNA based on homology or novel miRNAs based on small RNA sequencing reads using the previous pipeline. The new miRNA pipeline, mirMachine, was tested using The Arabidopsis Information Resource, TAIR10, release of the *Arabidopsis* genome and the International Wheat Genome Sequencing Consortium (IWGSC) wheat reference genome v2.

Introduction

Advances in next generation sequencing technologies have widened the understanding of RNA structures and regulatory elements, revealing functionally important non-coding RNAs (ncRNAs). Among different types of ncRNAs, microRNAs (miRNAs) constitute a fundamental regulatory class of small RNAs with a length between 19 and 24 nucleotides in plants[1,2]. Since the discovery of the first miRNA in the nematode *Caenorhabditis elegans*[3], the presence and the functions of miRNAs have been studied extensively in animal and plant genomes as well[4,5,6]. miRNAs function by

targeting mRNAs for cleavage or translational repression[7]. Accumulating evidence has also shown that miRNAs are involved in a wide range of biological processes in plants including growth and development[8], self-biogenesis[9], and several biotic and abiotic stress responses[10].

In plants, miRNAs are initially processed from long primary transcripts called pri-miRNAs[11]. These pri-miRNAs generated by RNA polymerase II inside the nucleus are long transcripts forming an imperfect fold-back structure[12]. The pri-miRNAs later undergo a cleavage process to produce endogenous single-stranded (ss) hairpin precursors of miRNAs called pre-miRNAs[11]. The pre-miRNA forms a hairpin-like structure wherein a single strand folds into a double-stranded structure to excise an miRNA duplex (miRNA/miRNA*)[13]. Dicer-like protein cuts both strands of the miRNA/miRNA* duplex, leaving 2-nucleotide 3'-overhangs[14,15]. The miRNA duplex is methylated inside the nucleus, which protects the 3'-end of the miRNA from degradation and uridylation activity[16,17]. A helicase unwinds the methylated miRNA duplex after export and exposes the mature miRNA to the RNA-induced silencing complex (RISC) in the cytosol[18]. One strand of the duplex is mature miRNA incorporated into RISC , whereas the other strand, miRNA*, is degraded. The miRNA-RISC complex binds to the target sequence leading to either mRNA degradation in case of full complementarity or translational repression in case of partial complementarity[13].

Based on the expression and biogenesis features, guidelines for miRNA annotation have been described[15,19]. With the defined guidelines, Lucas and Budak developed the SUmir pipeline to perform a homology-based *in silico* miRNA identification in plants[9]. The SUmir pipeline was composed of two scripts: SUmirFind and SUmirFold. SUmirFind performs similarity searches against known miRNA datasets through National Center for Biotechnology Information (NCBI) Basic Local Alignment Search tool (BLAST) screening with modified parameters to include hits with only 2 or fewer mismatches and to avoid bias towards shorter hits (blastn-short -ungapped -penalty -1 -reward 1). SUmirFold evaluates the secondary structure of the putative miRNA sequences from BLAST[20] results using UNAfold[21]. SUmirFold differentiates miRNAs from small interfering RNAs by the identification of the characteristics of hairpin structure. Moreover, it differentiates miRNAs from other ssRNAs such as tRNA and rRNA by the parameters, minimum fold energy index > 0.67 and GC content of 24-71%. This pipeline has been recently updated by adding two additional steps to (i) increase sensitivity, (ii) increase annotation accuracy, and (iii) provide genomic distribution of the predicted miRNA genes[22]. Given the high conservation of plant miRNA sequences[23], this pipeline was originally designed for homology-based miRNA prediction. Novel miRNAs, however, could not be accurately identified with this bioinformatics analysis as it heavily relied on sequence conservation of miRNAs between closely related species.

This paper presents a new and fully automated miRNA pipeline, mirMachine that 1) can identify known and novel miRNAs more accurately (for example, the pipeline now uses sRNA-seq-based novel miRNA predictions as well as homology-based miRNA identification) and 2) is fully automated and freely available. The outputs have also included the genomic distributions of the predicted miRNAs. mirMachine was tested for both homology-based and sRNA-seq-based predictions in wheat and *Arabidopsis* genomes. Although initially released as free software, UNAfold became a commercial software in the last decade. With this upgrade, the secondary structure prediction tool was switched from

UNAfold to RNAfold so that mirMachine can be freely available. Users can now execute a short submission script to run the fully automated mirMachine pipeline (examples are provided at **https://github.com/hbusra/mirMachine.git**).

Protocol

1. Software dependencies and installation

1. Install software dependencies from their home site or using conda.
    1. Download and install Perl, if it is not already installed, from its home site (https://www.perl.org/get.html).
    NOTE: Represented results were predicted using Perl v5.32.0.
    2. Download Blast+, an alignment program, from its home site (https://www.ncbi.nlm.nih.gov/books/NBK279671/) as an executable and as source code.
    NOTE: Represented results were predicted using the BLAST 2.6.0+.
    3. Install precompiled package of RNAfold from https://www.tbi.univie.ac.at/RNA/.
    4. Alternatively, install these softwares using the following conda: i) **conda install -c bioconda blast**; ii) **conda install -c bioconda viennarna**.

2. The mirMachine setup and testing

1. Download the latest version of the mirMachine scripts and the mirMachine submission script from GitHub, https://github.com/hbusra/mirMachine.git, and then set the scripts path into the **PATH**.
2. Use the test data provided at the GitHub to make sure that the mirMachine along with all its dependencies have been downloaded correctly.
3. Run the mirMachine on the test data shown below.
bash mirMachine_submit.sh -f iwgsc_v2_chr5A.fasta -i mature_high_conf_v22_1.fa.filtered.fasta -n 10
NOTE: Set the **-n** option to 10 as the test data contains only one chromosome of the wheat genome. At defaults, the **-n** option is set to 20.
4. Control the **hairpins.tbl.out.tbl** output files for the predicted mature miRNAs, their predicted precursors, and their locations on the chromosomes.
5. Check the log files for the program outputs and warnings.

3. Homology-based miRNA identification

1. Run the mirMachine using the bash script shown below:
bash mirMachine_submit.sh -f $genome_file -i $input_file -m $mismatches -n $number_of_hits
2. Check the predicted miRNAs. Find the output file named **$input_file.results.tbl.hairpins.tbl.out.tbl** for the predicted miRNAs. Find the output file named **$input_file.results.tbl.hairpins.fsa** for the pre-miRNA FASTA sequences. Find the output file named **$input_file.results.tbl.hairpins.log** for the hairpin log file.

4. Novel miRNA identification

1. Preprocess the sRNA-seq FASTQ files into proper FASTA format. Trim adaptors if needed. Do not trim low-quality reads; instead, remove them. Remove reads containing **N**. Convert the FASTQ file into FASTA file (**$input_file**).
2. Run the mirMachine using the bash script shown below.

bash mirMachine_submit.sh -f $genome_file -i $input_file -n $number_of_hits -sRNAseq -lmax $lmax -lmin $lmin -rpm $rpm

NOTE: **$mismatches** was set to 0 for sRNA-seq based predictions.

3. Check the predicted miRNAs. Find the output file named **$input_file.results.tbl.hairpins.tbl.out.tbl** for the predicted miRNAs. Find the output file named **$input_file.results.tbl.hairpins.fsa** for the pre-miRNA FASTA sequences. Find the output file named **$input_file.results.tbl.hairpins.log** for the hairpin log file.

5. Advance parameters

NOTE: The defaults are defined for all the parameters except for the genome file and the input miRNA file.

1. Set the **-db** option to a blast database to skip the building reference database within the pipeline.
2. Set the **-m** option to the number of mismatches allowed.
   NOTE: At defaults, **-m** option was set to 1 for homology-based predictions and 0 for the sRNA-seq-based predictions.
3. Set the **-n** to the number of hits to eliminate after alignment (default to 20). Change this based on the species.
4. Use the **-long** to assess the secondary structures for the suspect list.
5. Use the **-s** to activate the novel miRNA prediction based on sRNA-seq data.
6. Set the **-lmax** option to the maximum length of the sRNA-seq reads to include in the screening.
7. Set the **-lmax** option to the minimum length of the sRNA-seq reads to include in the screening.
8. Use the **-rpm** option to set the Reads Per Million (RPM) threshold.
   NOTE: For advanced parameters like the length of pri-miRNAs/pre-miRNAs, experienced users are encouraged to modify the scripts for their research of interest. Additionally, if the users intend to skip some steps or prefer to use modified outputs, the submission script can be modified by simply adding **#** at the beginning of the lines to skip those lines.

Representative Results

The miRNA pipeline, mirMachine, described above was applied to the test data for the fast evaluation of the performance of the pipeline. Only the high-confidence plant miRNAs deposited at miRBase v22.1 were screened against the chromosome 5A of IWGSC wheat RefSeq genome v2[24]. mirMachine_find returned 312 hits for the nonredundant list of 189 high-confidence miRNAs with a maximum of 1 mismatch allowed (**Table 1**). mirMachine_fold classified 49 of them as putative miRNAs depending on the secondary structure evaluation. The highest represented group of miRNAs was miR9666 with a total of 18 miRNAs identified (**Figure 1**). Some miRNAs shared the same mature miRNA, but processed from a different pre-miRNA sequence. These miRNAs were renamed by the miRNA family name followed by a unique number, e.g., miR156-5p-1 and miR156-5p-2. Among the 49 putative miRNAs, 20 non-redundant mature miRNA sequences were identified. Some miRNAs can be transcribed from more than one locus resulting in a higher number of miRNAs represented. In the test data, miR9666-3p-5 was represented twice: one on the sense strand (at 602887137) and the other

on the antisense strand (at 542053079). All locations are provided in the GitHub under the TestData output file named **mature_high_conf_v22_1.fa.filtered.fasta.results.tbl.hairpins.tbl.out.tbl**.

Expression evidence in one plant genome is sufficient, given the conservation of miRNAs in plants; however, a high-confidence miRNA dataset only provides a limited amount of data. Therefore, it is the user's preference to use the high-confidence and/or experimentally validated miRNAs as the reference dataset and skip the expression validation step, or to use all plant miRNAs available as the reference dataset and look for the expression evidence afterwards. Here, as the high-confidence miRNAs were used as the reference set, which had been validated experimentally in one of the plant genomes, the expression validation step was skipped for the test data.

mirMachine was benchmarked using monocot and dicot plants including *Arabidopsis thaliana* (*Arabidopsis*, TAIR10 release) and *Triticum aestivum* (wheat, IWGSC RefSeq v2). The performance of the homology-based and the sRNA-seq-based predictions was evaluated, and the results were compared with the miRDP2[25], an NGS-based miRNA prediction tool. Homology-based predictions were executed using the non-redundant list of plant mature miRNA sequences deposited at the miRbase v22[26]. sRNA-seq-based predictions were executed using the publicly available datasets; GSM2094927 for *Arabidopsis* and GSM1294661 for the wheat. In addition to raw results, the homology-based predictions were filtered for the expression evidence of mature miRNA and miRNA star sequences using the same sRNA-seq datasets.

Figure 2 shows the performance of each tool and the mirMachine settings on the two species. Sensitivity was calculated as the total number of known miRNAs identified divided by the total number of miRNAs identified. The results showed that mirMachine outperformed miRDP2 in terms of sensitivity and the true positive predictions in the *Arabidopsis* data. For the wheat data, mirMachine homology-based prediction, supported by expression evidence, provided better sensitivity than miRDP2. For both the genomes, miRDP2 predicted higher number of true positives compared to mirMachine sRNA-seq and homology-based predictions with expression evidence. It should be noted that miRDP2 lowers the expression threshold (RPM, reads per million) from 10 to 1 for the prediction of known miRNAs, resulting in higher true positive predictions. In general, the mirMachine can be used for the identification of both novel and known miRNAs. One advantage of the mirMachine is its ability to predict genome-wide distribution of the putative miRNAs without a limitation of specific tissues and conditions. Finally, the mirMachine is user-friendly and provides flexibility to adjust parameters such as number of hits, mismatches, length of miRNAs, and RPMs for specific research purposes. Taken together, the mirMachine provides accurate predictions for the putative miRNAs in the transcriptomes and the genomes of the plants.

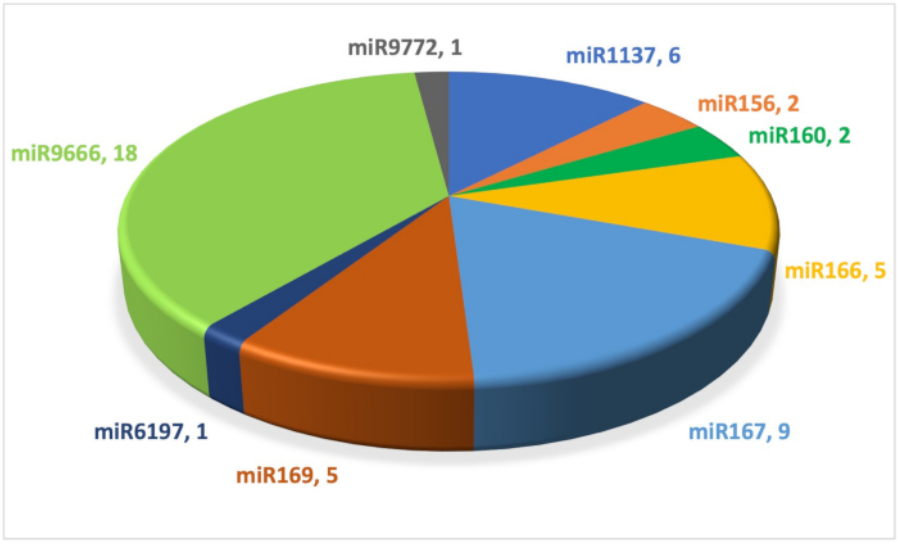

Figure 1: The distribution of miRNA families identified from the chromosome 5A of the IWGSC wheat reference genome v2. Data labels show the miRNA family and the number of miRNAs belonging to each miRNA family. Abbreviations: miRNA = microRNA; IWGSC = International Wheat Genome Sequencing Consortium. Please click here to view a larger version of this figure.

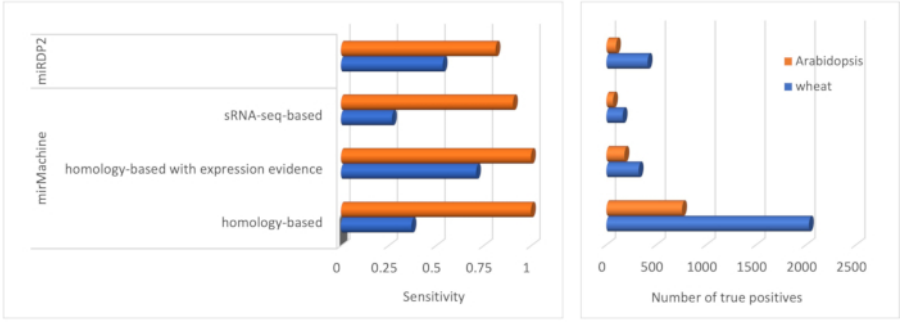

Figure 2: Performance assessment of the mirMachine. Comparisons of the sensitivity and the total number of known miRNAs predicted (true positives) are shown for the mirMachine with homology-based and sRNA-seq-based predictions and the miRDP2 software. Abbreviation: miRNA = microRNA. Please click here to view a larger version of this figure.

Genome	Genome Size	Reference miRNA dataset	mirMachine_find hits	mirMAchine_fold hits	# of miRNA families
Test data	~0.7 Gb	189	312	49	9
Chr5A					

Table 1: Statistics of the mirMachine. Test data are from the chromosome 5A of the IWGSC wheat reference genome v2. Abbreviations: miRNA = microRNA; IWGSC = International Wheat Genome Sequencing Consortium.

Discussion

Our miRNA pipeline, SUmir, has been used for the identification of many plant miRNAs for the last decade. Here, we developed a new, fully automated, and freely available miRNA identification and annotation pipeline, mirMachine. Furthermore, a number of miRNA identification pipelines including, but not limited to the previous pipeline, were dependent on UNAfold software[21], which became a commercial software over time, although once being freely available. This new and fully automated mirMachine is no longer dependent on the UNAfold; instead, the freely available RNAfold from the ViennaRNA package[27] is used for secondary structure prediction. Additionally, all scripts for the mirMachine were gathered in a bash script with adjustable parameters to make mirMachine a fully automated and freely available miRNA prediction and annotation tool.

The mirMachine benefited from the characteristics of plant miRNAs and their biogenesis. As opposed to animal pre-miRNAs, plant pre-miRNAs are variable in length and structural features[15]. Consequently, a criterion has been set for the identification of plant miRNAs depending on the characteristics of the miRNAs and their biogenesis[15]. No cut-off was set for the pre-miRNA length as the length of plant pre-miRNAs can vary remarkably and could be hundreds of nucleotides long. Instead, pri-miRNA structure folding, which was limited to ~700 bp in length, was first evaluated. Later, pre-miRNA sequence was predicted from the candidate pri-miRNA sequences and evaluated for proper folding statistics.

Many plant genomes, especially agronomically important cereals such as wheat and barley, possess highly repetitive genomes[28,29,30]. Other than the high-repeat content, polyploidy is observed in some of these plants[24], introducing additional complexities to the *in silico* identification and characterization of the miRNA structures. The repeats are a major source for the production of siRNAs[31], which resemble miRNAs in their mature forms; however, they differ in biogenesis and function[32,33]. It is extremely difficult to eliminate siRNAs from the candidate miRNA lists. In fact, the most widely used miRNA database, the miRBase[26], has been reported to contain large numbers of siRNAs annotated falsely as miRNAs[34,35]. Based on the differences in their biogenesis, the mirMachine filters the small RNAs that form a perfect pair with the antisense strand as siRNAs and places those sequences into the suspect table. Additionally, the mirMachine has the **-n** option, which defines the maximum number of hits to filter the candidate RNAs as siRNAs.

Expression evidence is required to validate all the miRNAs predicted *in silico*. As miRNAs are highly conserved among plant genomes, expression evidence in one of the plant genomes should be sufficient to confirm the validity of

the predicted miRNA. The use of high-confidence, mature miRNA sequences in the initial screening process has the advantage of providing expression evidence for all the predicted miRNAs; however, the short list of initial miRNA dataset limits the prediction of a comprehensive set of miRNAs in a genome. Alternatively, a full set of plant miRNAs deposited in the miRBase database can be used as an initial dataset instead of filtering for high-confidence miRNAs. Users are advised to look for expression evidence through expressed sequence tags, miRNA microarrays, or small RNA sequencing data for at least one of the plant genomes if any expression data are not available for the species of interest.

Homology-based miRNA predictions can help elucidate genome-wide distribution of the known family of miRNAs. These miRNAs are likely to be expressed in certain tissues and conditions. A drawback of homology-based predictions is the lack of ability to identify novel miRNA families. In contrast, sRNA-seq-based predictions could identify novel miRNAs with a cost of a high number of false positives. Therefore, the choice of the best approach is up to the users and the research of interest. The mirMachine presented here can help identification of the miRNAs based on either homology to known miRNAs or sRNA sequencing.

References

1. Voinnet, O. Origin, biogenesis, and activity of plant microRNAs. *Cell.* **136** (4), 669-687 (2009).
2. Budak, H., Akpinar, B. A. Plant miRNAs: biogenesis, organization and origins. *Functional & Integrative Genomics.* **15** (5), 523-531 (2015).
3. Lee, R. C., Feinbaum, R. L., Ambros, V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell.* **75** (5), 843-854 (1993).
4. Zhang, L. et al. Exogenous plant MIR168a specifically targets mammalian LDLRAP1: evidence of cross-kingdom regulation by microRNA. *Cell Research.* **22** (1), 107-126 (2012).
5. Pang, K. C., Frith, M. C., Mattick, J. S. Rapid evolution of noncoding RNAs: Lack of conservation does not mean lack of function. *Trends in Genetics.* **22** (1), 1-5 (2006).
6. Guleria, P., Mahajan, M., Bhardwaj, J., Yadav, S. K. Plant small RNAs: biogenesis, mode of action and their roles in abiotic stresses. *Genomics, Proteomics and Bioinformatics.* **9** (6), 183-199 (2011).
7. Jones-Rhoades, M. W., Bartel, D. P., Bartel, B. MicroRNAs and their regulatory roles in plants. *Annual Review of Plant Biology.* **57**, 19-53 (2006).
8. Singh, A. et al. Plant small RNAs: advancement in the understanding of biogenesis and role in plant development. *Planta.* **248** (3), 545-558 (2018).
9. Lucas, S. J., Budak, H. Sorting the wheat from the chaff: identifying miRNAs in genomic survey sequences of *Triticum aestivum* chromosome 1AL. *PloS One.* **7** (7), e40859 (2012).
10. Li, S., Castillo-González, C., Yu, B., Zhang, X. The functions of plant small RNAs in development and in stress responses. *Plant Journal.* **90** (4), 654-670 (2017).
11. Lee, Y., Jeon, K., Lee, J. T., Kim, S., Kim, V. N. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO Journal.* **21** (17), 4663-4670 (2002).
12. Lee, Y. et al. MicroRNA genes are transcribed by RNA polymerase II. *EMBO Journal.* **23** (2), 4051-4060 (2004).

13. Bartel, D. P. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell.* **116** (2), 281-297 (2004).
14. Lee, Y. et al. The nuclear RNase III Drosha initiates microRNA processing. *Nature.* **425** (6956), 415-419 (2003).
15. Meyers, B. C. et al. Criteria for annotation of plant microRNAs. *Plant Cell.* **20** (12), 3186-3190 (2008).
16. Sanei, M., Chen, X. Mechanisms of microRNA turnover. *Current Opinion in Plant Biology.* **27**, 199-206 (2015).
17. Li, J., Yang, Z., Yu, B., Liu, J., Chen, X. Methylation protects miRNAs and siRNAs from a 3′-end uridylation activity in Arabidopsis. *Current Biology.* **15** (16), 1501-1507 (2005).
18. Rogers, K., Chen, X. Biogenesis, turnover, and mode of action of plant microRNAs. *Plant Cell.* **25** (7), 2383-2399 (2013).
19. Axtell, M. J., Meyers, B. C. Revisiting criteria for plant microRNA annotation in the Era of big data. *Plant Cell.* **30** (2), 272-284 (2018).
20. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinformatics.* **10** (1), 421 (2009).
21. Markham, N. R. N., Zuker, M. UNAFold: Software for nucleic acid folding and hybridization. *Methods in Molecular Biology.* **453**, 3-31 (2008).
22. Alptekin, B., Akpinar, B. A., Budak, H. A comprehensive prescription for plant miRNA identification. *Frontiers in Plant Science.* **7**, 2058 (2017).
23. Zhang, B., Pan, X., Cannon, C. H., Cobb, G. P., Anderson, T. A. Conservation and divergence of plant microRNA genes. *Plant Journal.* **46** (2), 243-259 (2006).
24. Appels, R. et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science.* **361** (6403), eaar7191 (2018).
25. Wang, Y., Kuang, Z., Li, L., Yang, X. A bioinformatics pipeline to accurately and efficiently analyze the microRNA transcriptomes in plants. *Journal of Visualized Experiments: JoVE.* (155), e59864 (2020).
26. Kozomara, A., Griffiths-Jones, S. MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Research.* **42** (D1), D68-73 (2014).
27. Lorenz, R. et al. ViennaRNA Package 2.0. *Algorithms for Molecular Biology.* **6** (1), 26 (2011).
28. Wicker, T. et al. Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology.* 19 (1), 103 (2018).
29. Flavell, R. B., Bennett, M. D., Smith, J. B., Smith, D. B. Genome size and the proportion of repeated nucleotide sequence DNA in plants. *Biochemical Genetics.* **12** (4), 257-269 (1974).
30. Wicker, T. et al. The repetitive landscape of the 5100 Mbp barley genome. *Mobile DNA.* **8**, 22 (2017).
31. Yang, Q., Ye, Q.A., Liu, Y. Mechanism of siRNA production from repetitive DNA. *Genes and Development.* **29** (5), 526-537 (2015).
32. Lam, J. K. W., Chow, M. Y. T., Zhang, Y., Leung, S. W. S. siRNA versus miRNA as therapeutics for gene silencing. *Molecular Therapy. Nucleic Acids.* **4** (9), e252 (2015).
33. Bartel, B. MicroRNAs directing siRNA biogenesis. *Nature Structural and Molecular Biology.* **12** (7), 569-571 (2005).
34. Meng, Y., Shao, C., Wang, H., Chen, M. Are all the miRBase-registered microRNAs true? A structure-

and expression-based re-examination in plants. *RNA Biology.* **9** (3), 249-253 (2012).

35. Berezikov, E. et al. Evolutionary flux of canonical microRNAs and mirtrons in *Drosophila. Nature Genetics.* 42 (1), 6-9; author reply 9-10 (2010).